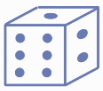
Probability is a good way to model uncertainty. We can build a model based on uncertainty.

$\Omega$ sample space contains all the possible events that might happen (atomic or not)
$w \in \Omega$ is an element of such space

$P: \Omega \to \mathbb{R}$ is a function such that $\begin{cases} 0 \leq P(w) \leq 1 \\ \sum_{w \in \Omega} P(w) = 1 \end{cases}$

e.g. rolling a die



$\Omega = \{1, 2, 3, 4, 5, 6\}$
$w_1, w_2, w_3, w_4, w_5, w_6$ atomic events $\in \Omega$


Before rolling a die there is uncertainty
$P(w) = \{1/6, 1/6, 1/6, 1/6, 1/6, 1/6\}$
after rolling a die there is no uncertainty

An event A is any subset of $\Omega$

$$P(A) = \sum_{w \in A} P(w)$$

We could also consider non-atomic events e.g. an event could be "the outcome is an even number" that correspond to

$$P(A) = \sum_{w \in A} P(w) = 1/6 + 1/6 + 1/6 = 3/6 = 1/2$$

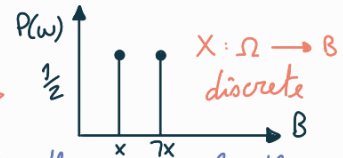
## Random variables

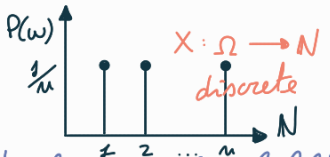
A random variable X is a variable whose possible values are numerical outcomes of a random phenomenon.
They are functions from the sample space to some range.
There are two types of random variables:

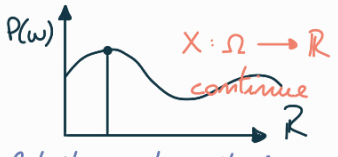
- discrete
- continue

"the outcome is $\leq 4$"

$$P(B) = \sum_{w \in B} P(w) = 1/6 + 1/6 + 1/6 + 1/6 = 4/6 = 2/3$$



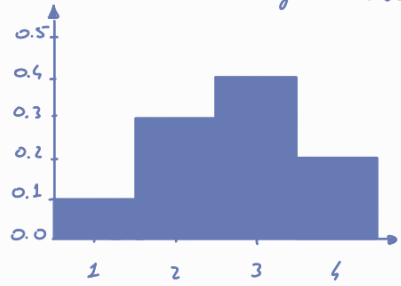

in these examples the events have equal probability but it is not said in general

A discrete random variable can take on only a countable number of distinct values such as 0, 1, 2, ...
If a random variable can take only a finite number of distinct values, then it must be discrete.
The probability distribution of a discrete random variable is a list of probabilities associated with each of its possible values. It is also called probability mass function (PMF)

Suppose a variable X can take the values 1, 2, 3, 4
The probabilities associated with each outcome are described by the table:

| Outcome | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| Probability | 0.1 | 0.3 | 0.4 | 0.2 |
| | $w_1$ | $w_2$ | $w_3$ | $w_4$ |

$$\sum_{w \in \Omega} P(w) = 0.1 + 0.3 + 0.4 + 0.2 = 1$$




probability histogram

the probability that X is equal to 2 or 3 is $P(X=2 \wedge X=3) = P(X=2) + P(X=3) = 0.3 + 0.4 = 0.7$

A continuous random variable is one which takes an infinite number of possible values.
Continuous random variables are usually measurements e.g. height, weight, ...
A continuous random variable is not defined at specific values. Instead it is defined over interval of values and it is represented by the area under a curve (integral). The probability of observing any single value is equal to 0 since the number of values which may be assumed by the random variable is infinite.
The curve has no negative value ($P(x) \geq 0 \ \forall x$) } density curve
The total area under the curve is = 1

gaussian curve is a density curve

If in a process any event has an equal probability of being observed, the curve describing the distribution is a rectangle and is called uniform distribution.


$X = x_i$ ⟶ the random variable $X$ has the value $x_i$
this is equivalent to $\{\omega \in \Omega \mid X(\omega) = x_i\}$

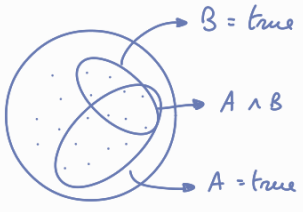
$$P(X = x_i) = P(\omega \in \Omega \mid X(\omega) = x_i) = \sum_{\{\omega \in \Omega \mid X(\omega) = x_i\}} P(\omega)$$

↳ atomic events that form the event $X = x_i$

**Propositions**: a proposition is the event (subset of $\Omega$) where an assignment to a random variable holds; propositions can be combined using standard logical operators

event $a$ = $A$ = true = $\{\omega \in \Omega \mid A(\omega) = \text{true}\}$
$\neg a$ = $A$ = false = $\{\omega \in \Omega \mid A(\omega) = \text{false}\}$
$a \wedge b$ = points where $A(\omega) = \text{true}$ and $B(\omega) = \text{true}$

$$P(a \wedge B) = \sum_{\{\omega \in \Omega \mid A(\omega) = a \wedge B(\omega) = b\}} P(\omega)$$




with propositions we are using multiple random variables defined on the same atomic event domain

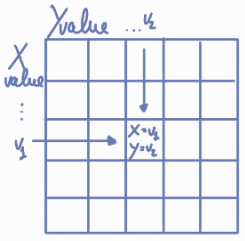
e.g. proposition on discrete random variables
$\Omega = \{\text{sunny}, \text{rain}, \text{cloudy}, \text{snow}\}$;
$\omega \in \Omega$ is an atomic event: it could happen that the weather is sunny / cloudy
$X : \Omega \to \mathbb{R}$ is a random variable
$X = \text{rain}$ is a proposition

**Prior probability** $P(h)$ denotes the initial (before analyzing the dataset) probability that hypothesis $h$ holds. It can reflect the background knowledge we have about the chance that $h$ is a correct hypothesis. If we have no such prior knowledge, then we might simply assign the same prior probability to each candidate hypothesis. $P(D)$ denotes the probability that training data $D$ will be observed.

Prior (or unconditional) probabilities of propositions correspond to belief prior to arrival of any (new) evidence.

$P(\text{Odd} = \text{true}) = 0.5$
$P(\text{Weather} = \text{sunny}) = 0.72$

**Joint probability distribution** for a set of random variables gives the probability of every atomic joint event on those random variables.

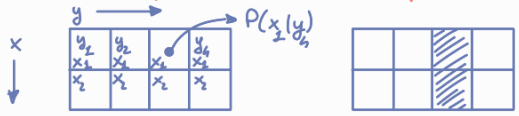



$P(X = \text{value}_1, Y = \text{value}_2)$

**Conditional / Posterior probability** $P(D \mid h)$ denotes the probability of observing data $D$ given that hypothesis $h$ holds. More generally, $P(x \mid y)$ denotes the probability of $x$ given $y$.
We are interested in $P(h \mid D)$: probability that $h$ holds given the observed training data $D$. $P(h \mid D)$ reflects the confidence that $h$ holds after we have seen the training data $D$.

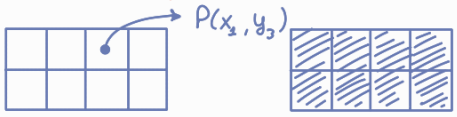
posterior ⟶ influence of the training data $D$ ⟶ knowledge
prior ⟶ independent of $D$
posterior ⟶ $P(\text{PlayTennis} = \text{true} \mid \text{Weather} = \text{sunny})$
prior ⟶ $P(\text{Weather} = \text{sunny})$

Posterior probabilities are different from joint probabilities and prior probabilities.




Conditional probability $P(x \mid y)$

$\sum = 1$




Joint probability $P(x, y)$

$\sum = 1$

P(PlayTennis | Weather) = 2 × 4 matrix

| P(PlayTennis \| Weather) | Sunny | Rain | Cloudy | Snow |
|---|---|---|---|---|
| PlayTennis = true | 0.8 | 0.2 | 0.8 | 0.1 |
| PlayTennis = false | 0.2 | 0.8 | 0.2 | 0.9 |
| | $\sum = 1$ | $\sum = 1$ | $\sum = 1$ | $\sum = 1$ |

Conditional probability $P(a|b)$ is defined as

$$P(a|b) \equiv \frac{P(a \wedge b)}{P(b)} \quad \text{if } P(b) \neq 0$$

denominator can be seen as a normalization constant $\alpha$

$$P(a|b) = \alpha P(a \wedge b)$$

from where we derive the product rule

$$P(a \wedge b) \equiv P(a|b)P(b) = P(b|a)P(a)$$

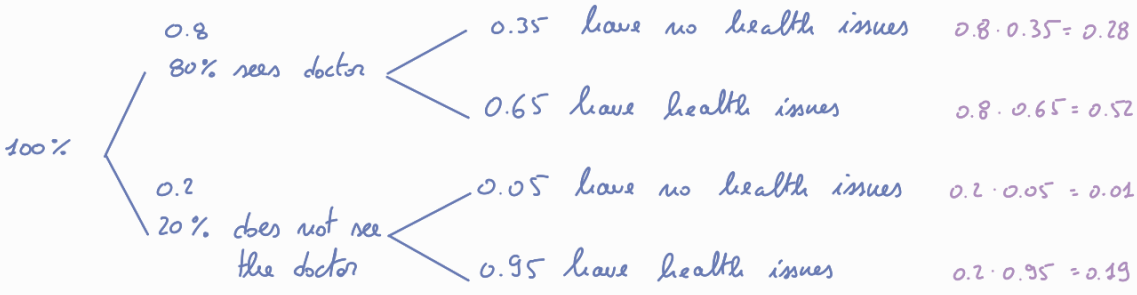
Total probabilities for a random variable $Y$ accepting mutually exclusive values $y_i$

$$P(X) = \sum_{y_i \in D(Y)} P(X|Y = y_i)\, P(Y = y_i)$$

$D(Y)$ : set of values for variable $Y$

e.g. 80% of people attend their primary care physician regularly; 35% of those people have no health problems during the following year. Out of 20% of people who don't see their doctor regularly, only 5% have no health issues during the following year. What is the probability a random person will have no health problems in the following year?

(Step 1) sketch out a tree; doing so, use information given in the question & obtained by the complement.




(Step 2) multiply the probabilities for each branch

(Step 3) find the probabilities that answer the question. The branches leading to "no health problems" are the top branch and the third branch.

$$0.28 + 0.01 = 0.29$$

$$P(X) = \sum_{y_i \in D(Y)} P(X|Y = y_i)\, P(Y = y_i) = \sum_{y_i \in D(Y)} P(X \wedge y_i) \quad \text{if the events } y_i \text{ are disjoint}$$

Chain rule is derived by successive application of product rule:

$$P(x_1, x_2) = P(x_1)\, P(x_2|x_1)$$

$$P(X_1, \ldots, X_n) = P(X_1, \ldots, X_{n-1}) P(X_n | X_1, \ldots, X_{n-1}) = \prod_{i=1}^{n} P(X_i | X_1, \ldots, X_{i-1})$$

**Inference by enumeration**: start from the joint distribution; for any proposition $\phi$, sum the atomic events where it is true

$$P(\phi) = \sum_{\omega : \omega \models \phi} P(\omega)$$

| | toothache | | ¬toothache | |
|---|---|---|---|---|
| | catch | ¬catch | catch | ¬catch |
| cavity | 0.108 | 0.012 | 0.072 | 0.008 |
| ¬cavity | 0.016 | 0.064 | 0.144 | 0.576 |

$\sum = 1$

| | toothache | | ¬toothache | |
|---|---|---|---|---|
| | catch | ¬catch | catch | ¬catch |
| cavity | 0.108 | 0.012 | 0.072 | 0.008 |
| ¬cavity | 0.016 | 0.064 | 0.144 | 0.576 |

$P(\text{toothache}) = 0.108 + 0.012 + 0.016 + 0.064 = 0.2$

| | toothache | | ¬toothache | |
|---|---|---|---|---|
| | catch | ¬catch | catch | ¬catch |
| cavity | 0.108 | 0.012 | 0.072 | 0.008 |
| ¬cavity | 0.016 | 0.064 | 0.144 | 0.576 |

$P(\text{cavity} \vee \text{toothache}) = 0.108 + 0.012 + 0.072 + 0.008 + 0.016 + 0.064 = 0.28$

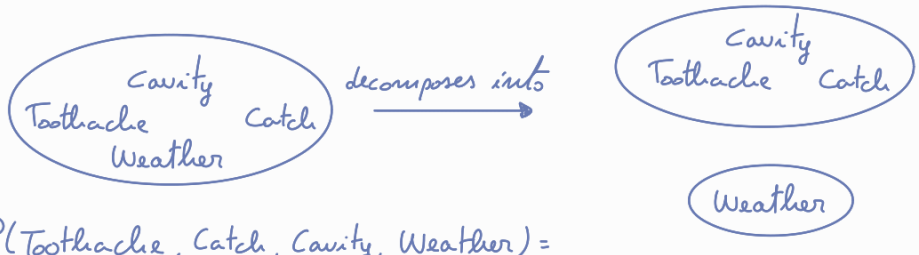
we can also compute conditional probabilities this way:

| | toothache | | ¬toothache | |
|---|---|---|---|---|
| | catch | ¬catch | catch | ¬catch |
| cavity | 0.108 | 0.012 | 0.072 | 0.008 |
| ¬cavity | 0.016 | 0.064 | 0.144 | 0.576 |

$$P(\neg\text{cavity} \mid \text{toothache}) = \frac{P(\neg\text{cavity} \wedge \text{toothache})}{P(\text{toothache})} = \frac{0.016 + 0.064}{0.108 + 0.012 + 0.016 + 0.064} = 0.4$$

**Independence**

a and b are *independent* iff $P(a|b) = P(a)$ or $P(b|a) = P(b)$ or $P(a,b) = P(a)P(b)$




$$P(\text{Toothache}, \text{Catch}, \text{Cavity}, \text{Weather}) = P(\text{Toothache}, \text{Catch}, \text{Cavity})\, P(\text{Weather})$$

**Conditional independence**

x and y are *conditionally independent* from z iff $P(x|y,z) = P(x|z)$

$$P(x,y|z) = P(x|y,z)P(y|z) = P(x|z)P(y|z)$$

with conditional independence, the chain rule becames:

$$P(y_1, \dots, y_m | z) = P(y_1|z)P(y_2|z) \dots P(y_m|z)$$

iff $y_i$ is conditionally independent from $y_j$ $\forall i,j \in [1, \dots, n]$

In most cases the use of the conditional independency hypothesis reduces the size of the representation of the joint distribution from exponential in n to linear in n; in general, it simplifies a lot the calculation, but it is always an assumption; complex systems have hundreds of variables, none of which are independent: conditional independence is a less strict condition.

# Bayes' Rule

Product rule ⟶ $P(a \wedge b) = P(a|b)P(b) = P(b|a)P(a)$

$$P(a|b) = \frac{P(b|a)P(a)}{P(b)} = \alpha P(b|a)P(a) \quad \text{in distribution form}$$

meaning:

$$P(\text{cause}|\text{effect}) = \frac{P(\text{effect}|\text{cause})P(\text{cause})}{P(\text{effect})}$$

Bayes rule and conditional independence

$$P(Z|Y_1, Y_2, \ldots, Y_n) = \alpha P(Y_1, Y_2, \ldots, Y_n|Z)P(Z)$$

meaning:

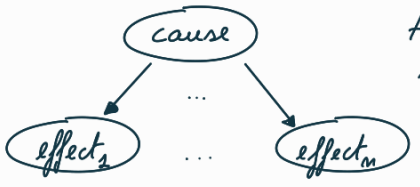
$$P(\text{cause}|\text{effect}_1, \ldots, \text{effect}_n) = \alpha P(\text{cause}) \prod_i P(\text{effect}_i|\text{cause})$$

conditional independence

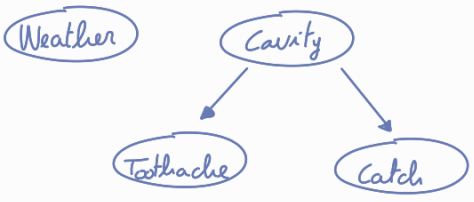
$$P(y_1, \ldots, y_m|z) = P(y_1|z)P(y_2|z)\ldots P(y_m|z)$$

the number of total parameters is linear in n

Bayesian networks are graphs with nodes denoting random variables and links denoting the dependencies between them.




A simple, graphical notation for conditional independence assertions and hence for compact specification of full joint distributions




Weather is independent of the other variables
Toothache and Catch are conditionally independent from each other given Cavity

Joint probabilities can be computed this way: $P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i|\text{Parents}(x_i))$

Burglar Bayesian Network Example:
I'm at work, neighbor John calls to say my alarm is ringing, but neighbor Mary doesn't call. Sometimes the alarm is set off by minor earthquakes. Is there a burglar?

Variables: Burglar (B), Earthquake (E), Alarm (A), JohnCalls (J), MaryCalls (M)

Network topology reflects causal knowledge:
- A burglar can set the alarm
- An earthquake can set the alarm
- The alarm can cause Mary to call
- The alarm can cause John to call